

OBJETIVO

Objetivo e problema

O objetivo deste trabalho é criar um pipeline de dados com foco na análise de uma base de dados sobre vendas de jogos de vídeo game. Esse pipeline permitirá a ingestão, processamento, transformação e análise dos dados. Com um pipeline bem definido, é possível melhorar a eficiência, precisão e agilidade para resolver problemas reais de negócios e na obtenção de insights valiosos sobre o mercado de jogos de vídeo game. Como por exemplo, saber quais os jogos mais vendidos, por região, por plataforma, como as vendas variaram ao longo dos anos etc. Dessa forma, quero por meio deste pipeline, entender o contexto do mercado de jogos de videogame e saber prioritariamente, **quais os 10 jogos mais vendidos no intervalo temporal desta base de dados**. Pretendo também responder as seguintes perguntas:

Como são as vendas anuais por região do mundo?

Qual estilo/gênero de jogo mais vendido no mundo?

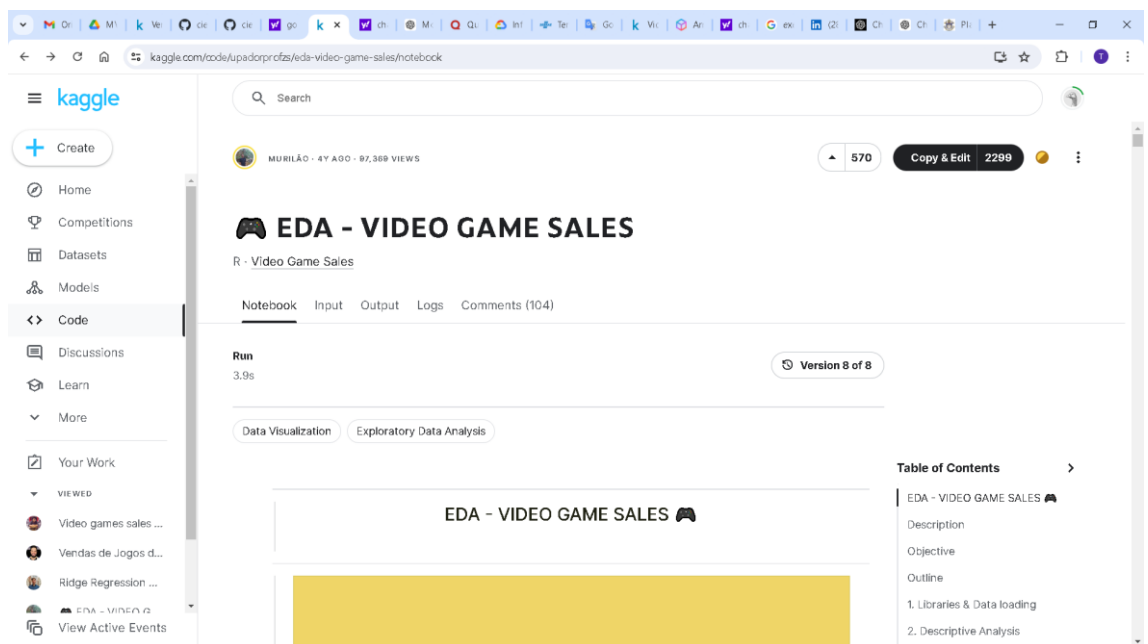
Qual Console é o campeão de vendas de jogos?

Quais Consoles com mais lançamentos?

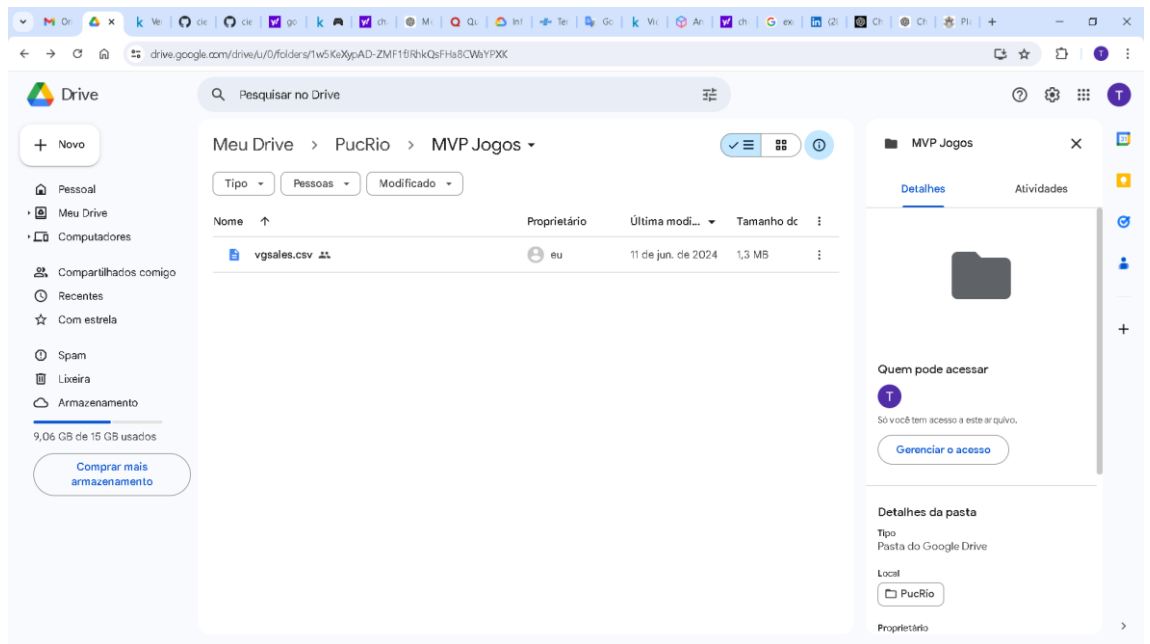
Qual o gênero de jogo mais vendido pelo console campeão de vendas?

Qual empresa publicadora é a campeã de vendas no mercado de jogos?

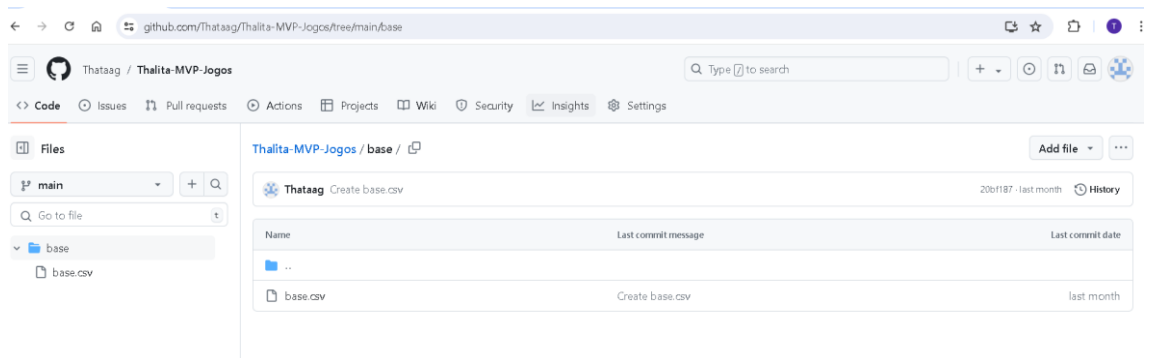
Peguei meu dataset no Kaggle:



Salvei no Drive:



Enviei pro GitHub e depois enviei pro Databricks:



MODELAGEM

Modelagem Flat Table

Flt mvp_jogos

Rank

Name

Platform

Year

Genre

Publisher

NA_Sales

EU_Sales

JP_Sales

Other_Sales

Global_Sales

O dataset foi baixado no Kaggle e é composto de 16.598 linhas e 11 colunas. Segue abaixo, o conteúdo de cada coluna:

Rank: Posição do jogo no ranking de vendas;

Name: Nome do jogo;

Platform: Console que lançou o jogo;

Year: Ano em que o jogo foi lançado;

Genre: Gênero/Estilo do jogo;

Publisher: Empresa responsável pela publicação do jogo;

NA_Sales: Vendas na América do Norte, em milhões de dólares;

EU_Sales: Vendas na Europa, em milhões de dólares;

JP_Sales: Vendas no Japão, em milhões de dólares;

Other_Sales: Vendas em outras regiões, em milhões de dólares;

Global_Sales: Venda total no mundo.

.....

As variáveis que pertencem à categoria quantitativa (numéricas) são: Rank, Year, NA_Sales, EU_Sales, JP_Sales, Other_Sales and Global_Sales.

As variáveis classificadas como qualitativas (categóricas) do conjunto de dados em estudo são: Name, Platform, Genre and Publisher.

.....

Domínios dos Dados Desejado:

- Rank: Valores inteiros representando a posição do jogo no ranking de vendas. **Valor mín: 1 e máx: 16600**

- Name: Strings não vazias com nomes dos jogos.

Resumo da coluna categórica 'Name':

Need for Speed: Most Wanted 12

Ratatouille 9

FIFA 14 9

LEGO Marvel Super Heroes 9

Madden NFL 07 9

...

Ar tonelico Qoga: Knell of Ar Ciel 1

Galaga: Destination Earth 1

Nintendo Presents: Crossword Collection 1

TrackMania: Build to Race 1

Know How 2 1

Name: count, **Length: 11493**.

- Platform: Strings representando plataformas (ex.: 'PS4', 'XOne', 'Switch').

Resumo da coluna categórica 'Platform':

DS 2163

PS2 2161

PS3 1329

Wii 1325

X360 1265

PSP 1213

PS 1196

PC 960

XB 824

GBA 822

GC 556

3DS 509

PSV 413

PS4 336

N64 319

SNES 239

XOne 213

SAT 173

WiiU 143

2600 133

NES 98

GB 98

DC 52

GEN 27

NG 12

SCD 6

WS 6

3DO 3

TG16 2

GG 1

PCFX 1 Name: count.

- Year: Valores inteiros representando o ano (ex.: 2019, 2020, 2021). **Valor mín:1980 e máx: 2020**

- Genre: Strings representando gêneros de jogos (ex.: 'Ação', 'Aventura', 'RPG').

Resumo da coluna categórica 'Genre':

Action 3316

Sports 2346

Misc 1739

Role-Playing 1488

Shooter 1310

Adventure 1286

Racing 1249

Platform 886

Simulation 867

Fighting 848

Strategy 681

Puzzle 582

Name: count.

- Publisher: Strings com nomes das editoras/Publicadoras (ex.: 'Nintendo', 'EA').

Resumo da coluna categórica 'Publisher':

Electronic Arts 1351

Activision 975

Namco Bandai Games 932

Ubisoft 921

Konami Digital Entertainment 832

...

Warp 1

New 1

Elite 1

Evolution Games 1

UIG Entertainment 1

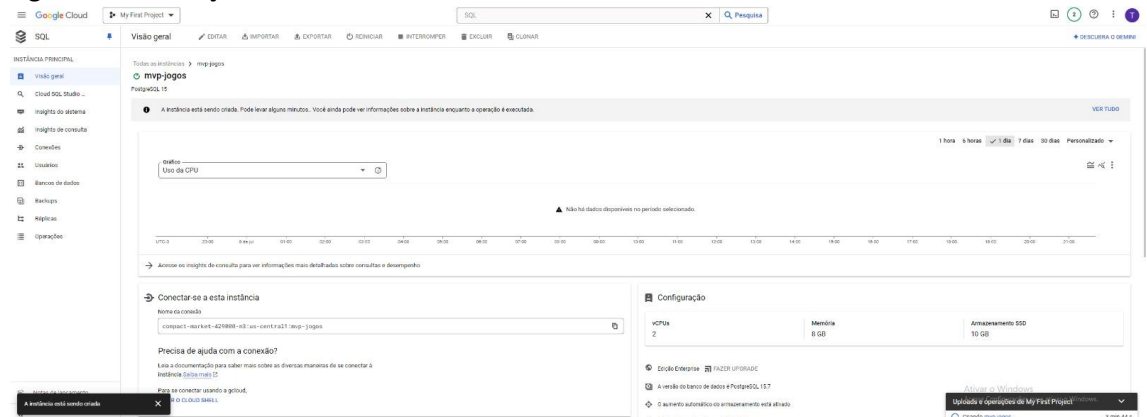
Name: count, **Length: 578.**

- NA_Sales: Valores decimais positivos representando as vendas na América do Norte. **Valor mín: 0 e máx: 41,49**

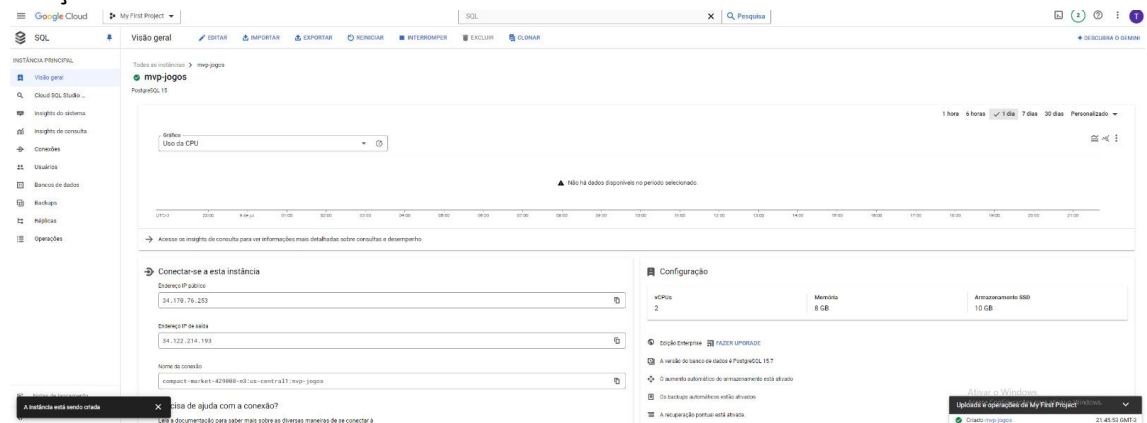
- EU_Sales: Valores decimais positivos representando as vendas na Europa. **Valor mín: 0 e máx: 29,02**
- JP_Sales: Valores decimais positivos representando as vendas no Japão. **Valor mín: 0 e máx: 10,22**
- Other_Sales: Valores decimais positivos representando as vendas em outras regiões. **Valor mín: 0 e máx: 10,57**
- Global_Sales: Valores decimais positivos representando as vendas globais, que é a soma das vendas regionais. **Valor mín: 0 e máx: 82,74**

CARGA

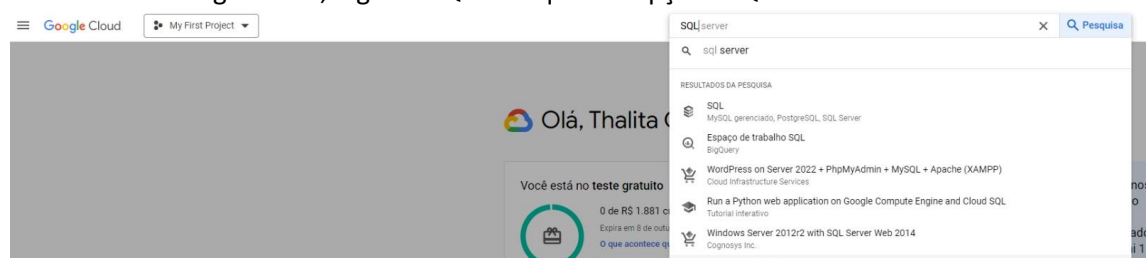
Aguardando criação do Banco do Dados:



Criação do Banco de Dados finalizada:



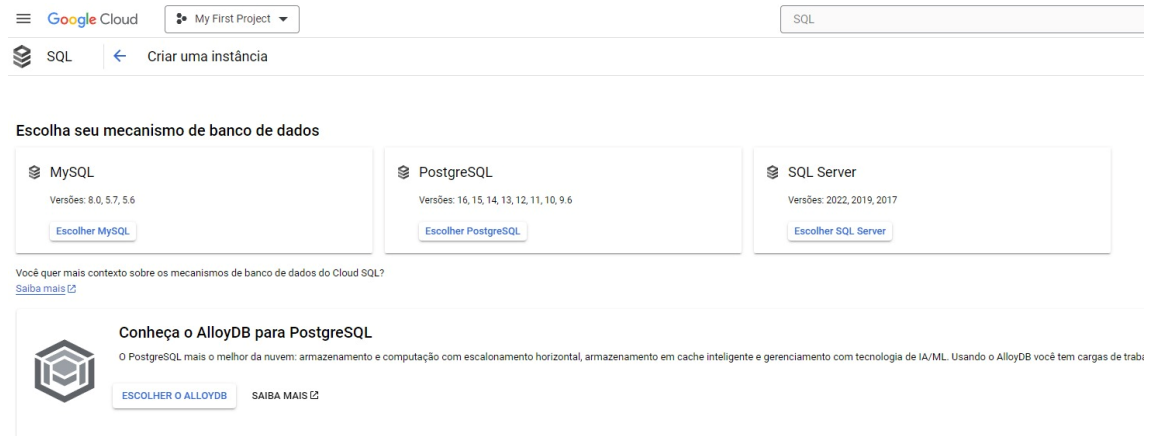
Na busca do Google cloud, digitei "SQL" e cliquei na opção "SQL":



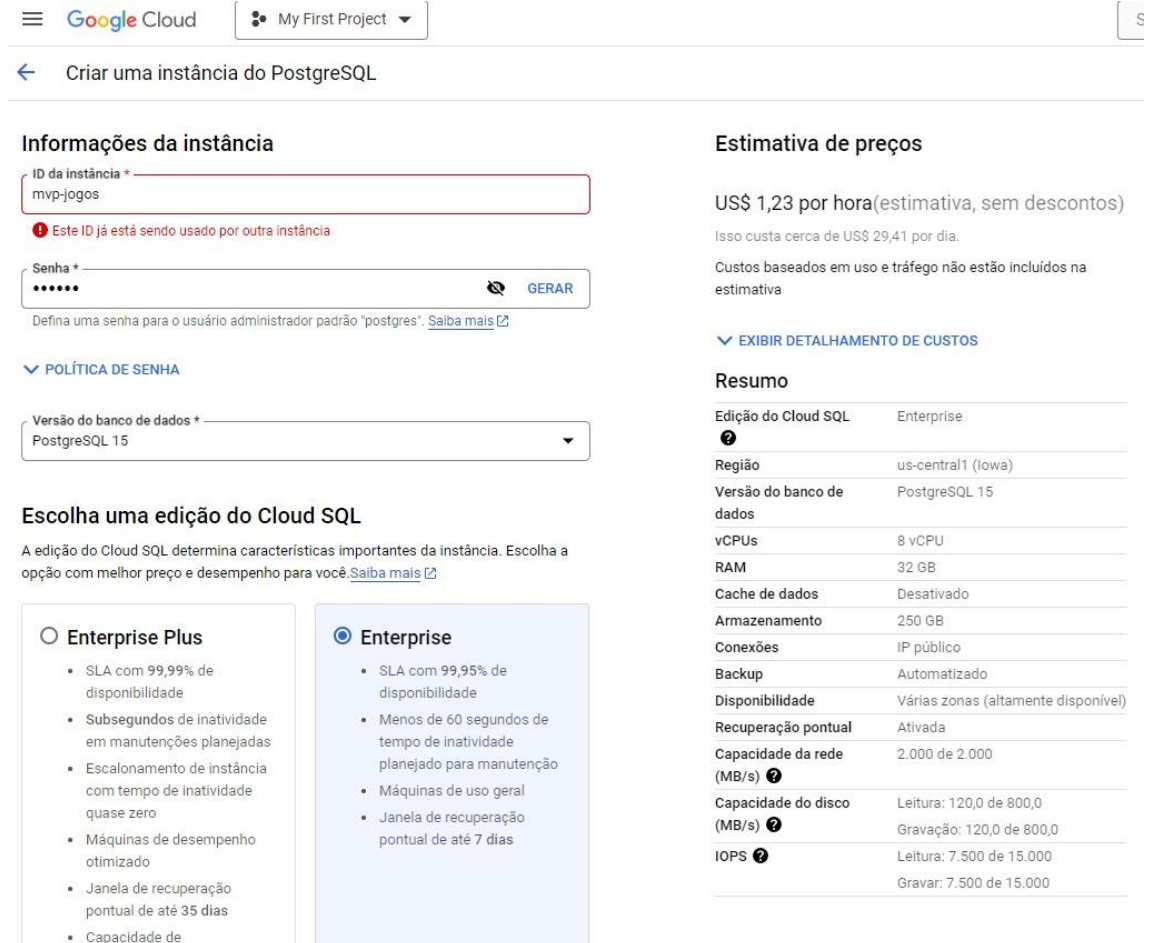
Na página do produto "SQL", cliquei em "CRIAR INSTÂNCIA":



Depois cliquei em PostgreSQL:



Coloquei o nome, senha, mantive a versão do PG, mudei a classe para Enterprise...:



Mudei a definição da máquina para "Sanbox", mantive as demais configurações e cliquei em "CRIAR INSTÂNCIA".:

Escolha uma predefinição para esta edição. Ela poderá ser personalizada depois se necessário.

Produção
8 vCPU, 32 GB de RAM, 250 GB de armazenamento, Altamente disponível

Desenvolvimento
4 vCPU, 16 GB de RAM, 100 GB de armazenamento, Única zona

Sandbox
2 vCPU, 8 GB de RAM, 10 GB de armazenamento, Única zona

Região

us-central1 (Iowa)

Disponibilidade por zona

☐ Única zona

Em caso de interrupção, não há failover. Não recomendado para produção.

☒ Várias zonas (altamente disponível)

Failover automático para outra zona na sua região selecionada. Recomendado para instâncias de produção. Aumenta o custo.

▼ ESPECIFICAR ZONAS

Personalizar sua instância

Também é possível personalizar as configurações da instância posteriormente

▼ MOSTRAR OPÇÕES DE CONFIGURAÇÃO

CRIAR INSTÂNCIA

CANCELAR

Para criar um banco de dados, vou em Banco de Dados e cliço em "CRIAR BANCO DE DADOS":

☰

Google Cloud

My First Project ▼

SQL

📌

INSTÂNCIA PRINCIPAL

📄

Visão geral

🔍

Cloud SQL Studio ...

📊

Insights do sistema

📈

Insights de consulta

🔗

Conexões

👤

Usuários

🗃️

Bancos de dados

💾

Backups

🔄

Réplicas

⋮

Operações

Bancos de dados

Todas as instâncias > mvp-jogos

✅ **mvp-jogos**

PostgreSQL 15

+

 CRIAR BANCO DE DADOS

Nome ↑	Compilação	Conjunto d
postgres	en_US.UTF8	UTF8

Dei o nome db e cliquei em CRIAR:

Criar um banco de dados

Nome do banco de dados *


db


Precisa seguir as regras do identificador do PostgreSQL. [Saiba mais](#)

CRIAR

CANCELAR


Para criar um usuário para o db, vim em usuários, cliquei em "ADICIONAR CONTA DE USUÁRIO":


 SQL





Usuários


INSTÂNCIA PRINCIPAL


 Visão geral


 Cloud SQL Studio ...


 Insights do sistema


 Insights de consulta


 Conexões

 **Usuários**


 Bancos de dados

 Backups

 Réplicas


 Operações




Todas as instâncias > mvp-jogos

 **mvp-jogos**

PostgreSQL 15

As contas de usuário permitem que os usuários e os aplicativos se conectem à sua instância. [Learn more](#)

 **ADICIONAR CONTA DE USUÁRIO**

	Nome de usuário ↑	Autenticação	Status da senha	
	postgres	Integrado	N/A	

Coloquei nome e senha e cliquei em "ADICIONAR":

Adicionar uma conta de usuário à instância mvp-jogos

Escolha como autenticar

Você pode gerenciar o acesso a esta instância usando a autenticação integrada do Cloud IAM ou do PostgreSQL. [Learn more](#)

☒ Autenticação integrada.

Cria um novo nome de usuário e senha específicos para esta instância. A conta do usuário tem acesso raiz `cloudsqlsuperuser`, mas isso pode ser personalizado mais tarde conforme necessário. [Learn more](#)

Nome de usuário *

Senha *



GERAR

Os usuários criados com autenticação integrada recebem o papel `cloudsqlsuperuser` e têm o mesmo conjunto de atributos que o usuário `postgres`. [Saiba mais](#)

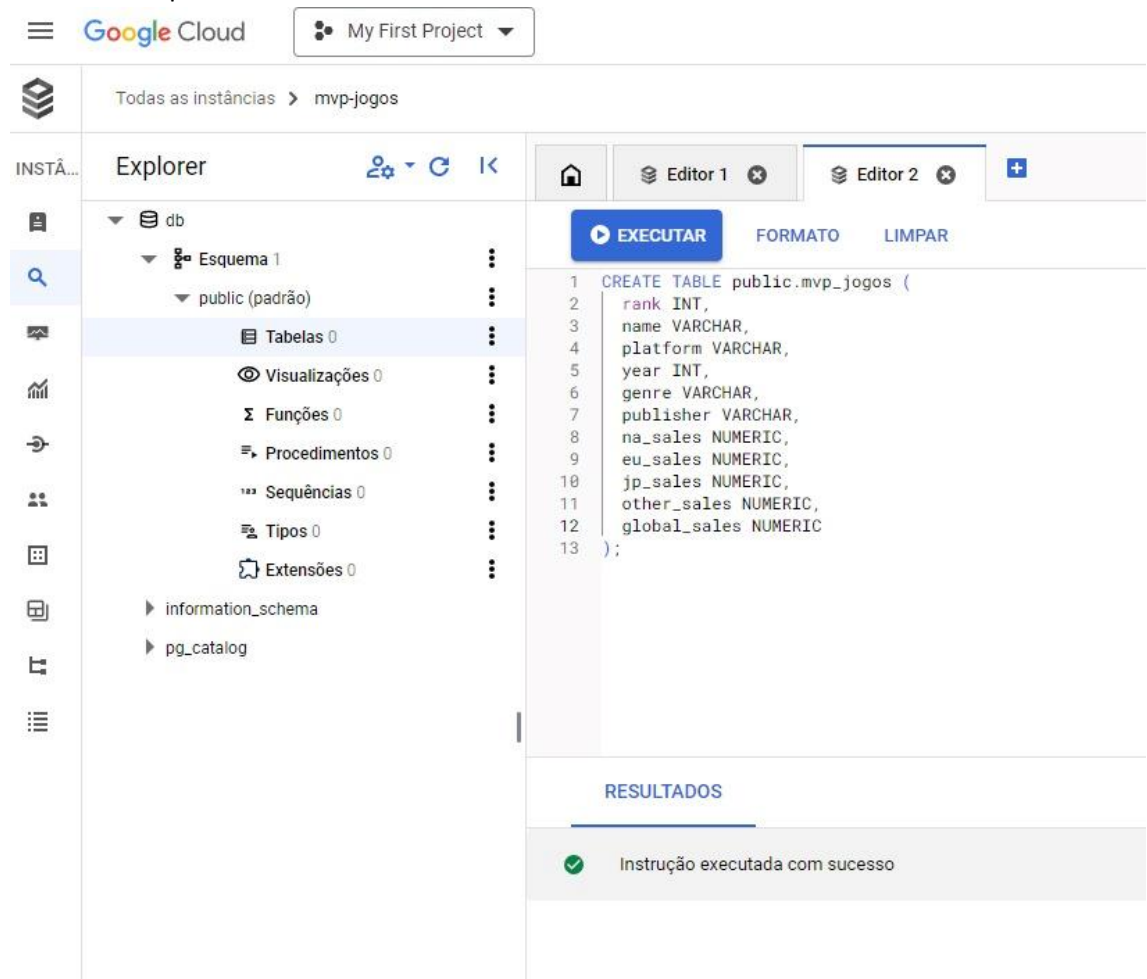
☐ Cloud IAM

Associa um participante atual IAM a esta conta de usuário. Para se conectar, é preciso ter um papel que fornece acesso à instância.

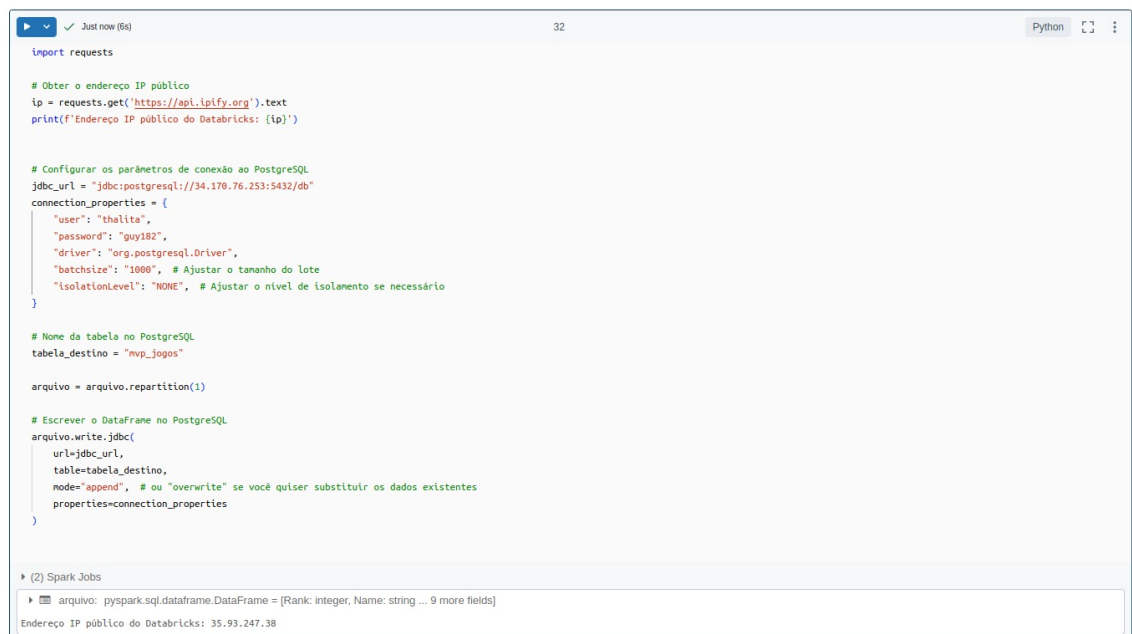
ADICIONAR

CANCELAR

Para criar a tabela na base, fui em "CLOUD SQL STUDIO", escrevi o script SQL para criar a tabela e cliquei em "EXECUTAR":



Sucesso na exportação de dados notebook para PG:



Contagem de registros exportados para o PG:



Status da avaliação gratuita: R\$ 1.868,12 de crédito e 90 dias restantes. Ative sua conta completa para ter acesso ilimitado



Google Cloud

My First Project

Pt



Todas as instâncias > mvp-jogos

INSTÂ...

Explorer



Editor 1



EXECUTAR

FORMATO

LIMPAR

```
1 select count(*) from mvp_jogos
2 |
```

RESULTADOS

count

16323

Visualização de registros exportados para o PG:



Status da avaliação gratuita: R\$ 1.868,12 de crédito e 90 dias restantes. Ative sua conta completa para ter acesso ilimitado a todos os recursos do Google Cloud. Use os créditos restantes e pague apenas pelo que usar.



Google Cloud

My First Project

Pesquise (/) recursos, documentos, produtos e muito mais

Pesquisa



Todas as instâncias > mvp-jogos

INSTÂ...

Explorer



Editor 1



EXECUTAR

FORMATO

LIMPAR

```
1 select * from mvp_jogos
2 |
```

RESULTADOS										
rank	name	platform	year	genre	publisher	na_sales	eu_sales	jp_sales	other_sales	global_sales
438	Sonic Rush	DS	2005	Platform	Sega	1.22	1.57	0.06	0.29	3.15
439	Resident Evil 6	PS3	2012	Shooter	Capcom	0.88	0.97	0.88	0.42	3.15
440	Halo: The Master Chief Collection	XOne	2014	Shooter	Microsoft Game Studios	1.89	0.99	0.03	0.24	3.15
441	FIFA Soccer 08	PS2	2007	Sports	Electronic Arts	0.68	0	0	2.46	3.14
442	Spider-Man	PS	2000	Action	Activision	1.7	1.25	0.02	0.16	3.13
443	Need for Speed III: Hot Pursuit	PS	1998	Racing	Electronic Arts	2.14	0.86	0	0.13	3.12
444	Star Wars Episode I Racer	N64	1999	Racing	Nintendo	2.31	0.62	0.14	0.04	3.12
445	Personal Trainer: Cooking	DS	2006	Misc	Nintendo	0.91	1	1.03	0.17	3.12
446	The Elder Scrolls IV: Oblivion	PS3	2007	Role-Playing	Ubisoft	1.69	0.87	0.14	0.42	3.12
447	Dragon Warrior IV	NES	1990	Role-Playing	Enix Corporation	0.08	0	3.03	0.01	3.12
448	Heavy Rain	PS3	2010	Adventure	Sony Computer Entertainment	1.29	1.27	0.06	0.5	3.12
449	Mass Effect 2	X360	2010	Role-Playing	Electronic Arts	1.99	0.82	0.03	0.27	3.11
450	FIFA 15	X360	2014	Sports	Electronic Arts	0.78	2.02	0	0.3	3.11
451	Dragon Ball Z: Budokai	PS2	2002	Fighting	Atari	2.17	0.28	0.55	0.08	3.09
452	Madden NFL 2002	PS2	2001	Sports	Electronic Arts	2.5	0.16	0.01	0.42	3.08
453	World Class Track Meet	NES	1986	Sports	Namco Bandai Games	1.92	0.45	0.64	0.07	3.08
454	Borderlands 2	X360	2012	Shooter	Take-Two Interactive	1.89	0.91	0.04	0.24	3.07
455	The Legend of Zelda: A Link Between Worlds	3DS	2013	Action	Nintendo	1.4	0.99	0.46	0.22	3.07
456	Donkey Kong	GB	1994	Platform	Nintendo	1.57	0.62	0.55	0.34	3.07
457	The Sims: Vacation	PC	2002	Simulation	Electronic Arts	1.72	1.21	0	0.14	3.07

AUTOAVALIAÇÃO

Considero que meu trabalho foi simples e essa atividade foi bastante desafiadora pra mim, tendo em vista que essa foi minha primeira pipeline de dados e primeiro contato com o Databricks e com Engenharia de dados, contudo, acredito que cumpri o que foi proposto como atividade final do módulo.

Optei por utilizar a linguagem python porque me dá mais segurança.

O trabalho acima foi realizado no intuito de construir um pipeline de dados a partir de um dataset sobre vendas de videogames durante o período de 1980 à 2016, é um tema que me atrai e considero interessante.

Baixei do Kaggle e após a extração e carga no Databricks, explorei a base para entender sua composição e produzir o catálogo de dados/modelagem. Em seguida comecei a tratar os dados. No processo de transformação, foi identificado os registros nulos nas tabelas "Year" e "Publisher", onde optei por excluir dados cujo ano não foi informado e classifiquei como unknown onde não havia registro na coluna "Publisher".

Altere o tipo da coluna "Year" de string para int.

Analisei o período do dataset e observei que os anos de 2020 e 2017 apresentavam dados faltantes e os anos de 2018 e 2019 não apareciam na lista, então considerei que os lançamentos após o ano de 2016 eram incompletos/ruidosos e poderiam atrapalhar este estudo. Dessa forma, optei por remover estes lançamentos do dataset, definindo o período de 1980 a 2016. Nessa etapa, observei também que, a coluna "Rank" não estava considerando os jogos multiplataforma, ou seja, aqueles jogos que foram comercializados para e por diferentes consoles. Para contabilizar a venda dos jogos multiplataforma e visualizá-los num rank, somei as vendas totais desses jogos em cada console que o comercializou. Após isso, pude visualizar os dez jogos mais vendidos entre 1980 e 2016 em resposta ao meu problema e principal pergunta a ser respondida neste estudo.

Fiz a carga no postgresSQL, apenas para cumprir esse requisito, mas tanto a etapa de transformação, quanto a de análise, fiz no próprio Databricks.

Após isso fiz uma análise sobre a qualidade dos dados e, finalizei com as análises e proporcionaram respostas à todas as demais perguntas do problema além de mais algumas que surgiram ao longo da análise e confabulações a respeito do mercado de jogos de videogames, descritas na conclusão e solução do problema.

CONCLUSÃO e Solução do problema:

Após as etapas de ETL, pude deixar a base em condições para que eu pudesse realizar as análises que responderiam minhas perguntas e solução do meu problema. Por meio deste trabalho, foi possível extrair informações relevantes sobre a indústria de jogos eletrônicos no mundo e sobre as tendências e padrões do mercado global de videogames. Ao examinar os dados, conseguimos identificar várias características importantes que ajudam a compreender melhor a dinâmica deste setor.

Os 10 jogos mais vendidos considerando os jogos multiplataforma foram:

Wii Sports

Grand Theft Auto V

Super Mario Bros.

Tetris

Mario Kart Wii

Wii Sports Resort

Pokemon Red/Pokem...

Call of Duty: Mod...

New Super Mario B...

Call of Duty: Bla...

Vendas por Região: As regiões da América do Norte, Europa, Japão e outras partes do mundo apresentaram padrões distintos de vendas ao longo dos anos. A América do Norte e a Europa dominaram as vendas globais.

Gêneros de jogos mais vendidos no mundo: Gêneros como ação, esporte e tiro lideraram as vendas globais. Esta análise ajudou a destacar quais tipos de jogos têm maior apelo no mercado.

Console campeão de vendas: PS2.

Os 10 jogos mais vendidos de ps2 foram:

Grand Theft Auto

Grand Theft Auto II

Gran Turismo 3

Grand Theft Auto III

Gran Turismo 4

Final Fantasy X

Need for Speed

Need for Speed

Medal of Honor

Kingdom Hearts

E os gêneros de jogos mais vendidos de PS2 foram:

Sports

Action

Racing

Shooter

Misc

Role-Playing

Fighting

Platform

Simulation

Adventure

Strategy

Puzzle

Console com mais lançamentos: DS.

As plataformas que mais venderam jogos ao longo dos anos foram identificadas, com destaque para consoles icônicos como o PlayStation 2, Xbox 360 e Nintendo Wii. Estes consoles não apenas tiveram um alto volume de vendas, mas também um número significativo de lançamentos. A análise dos lançamentos por plataforma revelou que certas plataformas têm um número significativamente maior de jogos lançados, o que pode influenciar a popularidade e as vendas desses consoles.

Publicadora campeã de vendas: Nintendo.

Publicadora com mais lançamentos: Electronic Arts

As publicadoras que mais venderam jogos, como Nintendo, Electronic Arts e Activision, foram identificadas, demonstrando a importância dessas empresas no mercado global

de videogames. Elas não apenas lideram em vendas, mas também têm um grande número de lançamentos.

Implicações e Recomendações

- **Estratégias de Marketing Regionalizadas:** As diferenças nas vendas por região sugerem que as estratégias de marketing e distribuição devem ser adaptadas para atender às preferências regionais. Campanhas focadas em gêneros populares em cada região podem aumentar a eficácia.
- **Foco em Consoles e Gêneros Populares:** Investir no desenvolvimento de jogos para as plataformas mais vendidas e nos gêneros mais populares pode maximizar o retorno sobre o investimento para desenvolvedores e publicadoras.
- **Análise de Tendências Futuras:** Continuar a monitorar as tendências de vendas pode ajudar a prever mudanças no mercado e ajustar as estratégias de desenvolvimento e marketing conforme necessário.

Limitações e Trabalho Futuro

Embora a análise tenha fornecido insights valiosos, algumas limitações devem ser consideradas. A base de dados pode não incluir todas as vendas ou lançamentos, e os dados históricos podem não refletir completamente as tendências atuais. Trabalhos futuros podem incluir a análise de dados adicionais, como vendas digitais e mobile, para obter uma visão mais abrangente do mercado.

Em resumo, a análise de dados de vendas de videogames oferece uma compreensão aprofundada do mercado global, ajudando desenvolvedores, publicadoras e estrategistas a tomar decisões informadas para alcançar o sucesso no competitivo mundo dos videogames.